dream: Powerful differential expression analysis for repeated measures designs

Gabriel E. Hoffman*[1,2,3], Panos Roussos[1,2,3,4,5]

[1] Pamela Sklar Division of Psychiatric Genomics,
[2] Icahn Institute for Genomics and Multiscale Biology,
[3] Department of Genetics and Genomic Sciences,
[4] Department of Psychiatry,
Icahn School of Medicine at Mount Sinai, New York, NY, USA
[5] Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, NY, USA

* Corresponding author: G.E. Hoffman (gabriel.hoffman@mssm.edu)

Large-scale transcriptome studies with multiple samples per individual are widely used to study disease biology. Yet current methods for differential expression are inadequate for these repeated measures designs. Here we introduce a novel method, dream, that increases power, controls the false positive rate, integrates with standard workflows, and yields biological insight in multiple datasets.

Recent advances in the scale of transcriptomic and, more generally, functional genomic studies has enabled assaying individuals from multiple tissues[1], brain regions[2], cell types[3], time points[4–6] or induced pluripotent stem (iPS) cell lines[7–10]. These studies with multiple samples from each individual can test region- or context-specific effects, and decouple biological from technical variation in gene expression. Yet current analysis methods do not adequately model the complexity of these studies and can result in loss of power or, more problematically, a large number of false positive findings[11,12].

While other software gives state-of-the-art performance on datasets with a single measurement per individual[13–16], these methods are not adequate for repeated measures designs. This is not a shortcoming of a particular normalization procedure, but rather due to the common assumption that samples are statistically independent. This core assumption is violated in the case of repeated measures and existing methods can perform very poorly in this case. The challenge of modeling repeated measures, and especially the risk of false positives, has been raised[11,12]. Yet there is currently no adequate statistical solution for differential analysis of transcriptome or functional genomics data.

Here we present a novel statistical model, dream, that enables powerful analysis of repeated measures data while properly controlling the false positive rate. Dream is available within the variancePartition[17] Bioconductor package (`http://bioconductor.org/packages/variancePartition/`) and combines:

- random effects estimated separately for each gene[17]
- ability to model multiple random effects[18]
- fast hypothesis testing for fixed effects in linear mixed models[19]
- small sample size hypothesis test to increase power[20]
- empirical Bayes moderated t-test[21]
- precision weights to model measurement error in RNA-seq counts[13,22]
- seamless integration with the widely used workflow of limma[23]

The performance of dream was compared to current methods on biologically realistic simulations. The methods can be divided into four categories: 1) dream using default settings or a Kenward-Roger approximation (termed dream-KR), which is more powerful but much more computationally demanding; 2) duplicateCorrelation from the limma/voom workflow[23]; 3) DESeq2[14] and limma/voom[13] including all samples but ignoring the repeated measures design; and 4) DESeq2 and limma/voom with only a single replicate per individual.

The two dream methods are more powerful than the other methods (**Fig. 1**). Across a range of simulations of 4 to 50 individuals each with 2 to 4 biological replicates, the two dream methods have a lower false discovery rate (**Fig. 1A, Supplementary Fig. 1**), better precision-recall curves (**Fig. 1B, Supplementary Fig. 2**), and larger area under the precision-recall (AUPR) curve (**Fig. 1C, Supplementary Fig. 3**).

A test of differential expression must control the false positive rate accurately in order to be useful in practice. As expected[11,12], the methods that include all samples but ignore the correlation structure do not control the false positive rate (**Fig. 1D**). This lack of type I error control is present in all simulation conditions (**Fig. 1E, Supplementary Fig. 4**). Even more concerning, increasing the number of repeated measures can dramatically increase the false positive rate. Notably, duplicationCorrelation shows a slight increase in type I error at larger sample sizes. Higher type I error can translate into hundreds of false positive differentially expressed genes even when no genes are truly differently expressed (**Supplementary Fig. 5**). Importantly, both versions of dream accurately control the type I error with sufficient sample size.

The two versions of dream give the highest AUPR across all simulation conditions (**Fig. 1F**) while properly controlling the false positive rate. While dream-KR gives the best performance, especially at small sample sizes, the computational time required can be prohibitive. Using dream with the default settings performs nearly as well in simulations, but can be 2-20x faster (**Supplementary Fig. 6**).

Applying dream to empirical data gives biological insight for 3 neuropsychiatric diseases with different genetic architectures (**Fig. 2**). Alzheimers's disease is a common neurodegenerative disorder with a complex genetic architecture[24]. In analysis of RNA-seq data from 4 regions of post mortem brains from 26 individuals[25], dream identified known patterns of dysregulation in genes involved in adipogenesis, inflammation and monocyte response associated with Braak stage, a neuropathological metric of disease progression (**Fig. 2A, Supplementary Fig. 7**). Applying duplicateCorrelation only recovered a subset of these findings and produced larger false discovery rates across many biologically relevant gene sets. In order to avoid using arbitrary cutoffs to identify differentially expressed genes, gene set enrichments were evaluated using the differential expression t-statistics from each analysis (see Online Methods). Notably, the difference between dream and duplicateCorrelation is due to the way that these methods account for expression variation explained by variance across individuals (**Fig. 2B**). Genes with expression variation across individuals that is larger than the genome-wide average are susceptible to being called as false positive differentially expressed genes (see Online Methods) (**Fig. 2C**).

Timothy syndrome is a monogenic neurodevelopmental disorder caused by variants in the calcium channel CACNA1C. Induced pluripotent and derived cell types were generated from 2 affected and 4 unaffected individuals an expression was assayed by microarray[26]. Since up to 6 lines were generated per donor for each cell type, it is necessary to account for repeated measures design. Analysis with dream identified differentially expressed genes enriched for brain, neuron, synapse and ion channel function, while duplicateCorrelation was not able to identify many of these gene set enrichments (**Fig. 2D, Supplementary Fig. 8**).

Childhood onset schizophrenia is a severe neurodevelopment disorder, but the genetic cause is complex with patients having a higher rate of schizophrenia-associated copy number variants, as well as higher schizophrenia polygenic risk scores[27,28]. RNA-seq data was generated from iPS-derived neurons and neural progenitor cells from 11 patients with childhood onset schizophrenia[7] and 11 controls with up to 3 lines per donor and cell type. Analysis with dream identified gene sets involved in neuronal function at the 5% and 1% FDR levels that were not identified by duplicateCorrelation (**Supplementary Fig. 9, 10**).

We have demonstrated that dream has superior performance in biologically realistic simulations while retaining control of the false positive rate. Furthermore, dream is able to identify biologically meaningful gene set enrichments in three neuropsychiatric disorders with different genetic architectures where the current standard for repeated measures designs, duplicateCorrelation, cannot. Since dream is built on top of the limma[23] and variancePartition[17] workflow, it can easily accommodate expression quantifications from multiple software packages including featureCounts[29], kallisto[30], salmon[31], and RSEM[32], among others. Moreover, dream works seamlessly for differential analysis of ATAC-seq or histone modification ChIP-seq data. The power, type I error control, simple R interface, speed and flexibility of dream enables analysis of transcriptome and functional genomics data with repeated measures designs.

Online Methods

Linear mixed models are commonly applied in biostatistics, in order to account for the correlation between observations from the same individual in a repeated measures study[33,34]. We start with a description of a simple linear model for differential expression analysis and build towards the dream model.

Linear models for differential expression

Consider a linear model for a single gene

$$y_g = X\beta_g + \varepsilon_g \tag{1}$$

where y_g is a vector of $\log_2$ counts per million for gene g, the matrix X stores covariates as columns, β_g is the vector of regression coefficients, and ε_g is normally distributed error. In order to account for heteroskedastic error from RNA-seq counts, the error takes the form

$$\varepsilon_g \sim \mathcal{N}(0, diag(w_g)\sigma^2_g) \tag{2}$$

where σ^2_g is the residual variance, and w_g is a vector of precision weights[13]. Precision weights can be learned from the data in order to account for counting error in RNA-seq or variation in sample quality[13,22]. In this case, the estimates $\hat{\beta}_g$ can be obtained by a closed form least squares model fit.

Moderated t-statistics for linear models

Since this model is fit for thousands of genes, the widely used limma model[21] uses a hierarchical model that imposes a prior distribution on the residual variances σ^2_g. This approach borrows information across genes and produces moderated t-statistics that increase power and reduce false positives compared to standard t-statistics. Briefly, the moderated t-statistic for gene g and a subset of regression coefficients k is defined by a function f where

$$\begin{aligned}
\tilde{t}_{g,k} &= f_{d_0,\sigma^2_0}\left(\hat{\beta}_{g,k}, \hat{\sigma}^2_g, V_g, d\right) && (3)\\
&= \sqrt{\frac{d_0+d}{d}}\frac{\hat{\beta}_{g,k}}{\sqrt{\left[\hat{\sigma}^2_g + (d_0/d)\sigma^2_0\right] V_{g,k}}} && (4)
\end{aligned}$$

and follows a t-distribution with $d + d_0$ degrees of freedom, where $\hat{\beta}_{g,k}$ and $\hat{\sigma}^2_g$ are the estimates of the regression coefficients and residual variance, respectively; $V_{g,k}$ is the unscaled covariance matrix of $\hat{\beta}_{g,k}$; and d is the residual degrees of freedom of the model fit. The terms d_0 and σ^2_0 are the prior residual degrees of freedom and prior residual variance, respectively. These values are estimated using an empirical Bayes approach combining the models fit for all genes. See Smyth[21] for details of the empirical Bayes estimation of d_0 and σ^2_0, and derivation of f.

Accounting for repeated measures with a two step model: duplicate correlation

The most widely used approach for handing repeated measures in differential expression analysis is the `duplicateCorrelation()` function available in limma[23]. This approach involves two steps. In the first step, a linear mixed model is fit for each gene separately, and only allows a single random effect. The model is

$$\begin{aligned} y_g &= X\beta_g + Z\alpha_g + \varepsilon_g && (5) \\ \alpha_g &\sim \mathcal{N}(0, \tau_g^2) && (6) \end{aligned}$$

where Z is the design matrix for the random effect, with coefficients α_g drawn from a normal distribution with variance τ_g^2. After fitting this model for each gene, a single genome-wide variance term is computed according to

$$\tau^2 = tanh\left(\frac{1}{G}\sum_{g=1}^{G} atanh\left(\tau_g^2\right)\right) \tag{7}$$

where G is the number of genes, $tanh$ is the hyperbolic tangent and $atanh$ is its inverse.

In the next step, this single variance term, τ^2, is then used in a generalized least squares model fit for each gene, blocking by individual:

$$\begin{aligned} y_g &= X\beta_g + \varepsilon_g && (8) \\ \varepsilon_g &\sim \mathcal{N}(0, diag(w_g)\Sigma_\varepsilon) && (9) \\ \Sigma_\varepsilon &= \begin{pmatrix} 1 & \tau^2 & 0 & 0 & 0 \\ \tau^2 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 1 & \tau^2 \\ 0 & 0 & 0 & \tau^2 & 1 \end{pmatrix} && (10) \end{aligned}$$

where Σ_ε is the covariance between samples and considers the correlation between samples from the same individual. Note that the same τ^2 value is used for all genes. The moderated t-statistics are defined as before.

The duplicateCorrelation method allows the user to specify a single random effect usually corresponding to donor. So it can't model multi-level design. Moreover, duplicateCorrelation estimates a single variance term genome-wide even though the donor contribution of a particular gene can vary substantially from the genome-wide trend[17]. Using a single value genome-wide for the within-donor variance can reduce power and increase the false positive rate in a particular, reproducible way. Consider the variance component for gene g, τ_g^2, compared to the single genome-wide value, τ^2. For genes where $\tau_g^2 > \tau^2$, using τ^2 under-corrects for the donor component so that it increases the false positive rate of gene g compared to using τ_g^2. Conversely, for genes where $\tau_g^2 < \tau^2$, using τ^2 over-corrects for the donor component so that it decreases power for gene g. Increasing sample size does not overcome this issue.

Using the single variance term genome-wide and using the *tanh* and *atanh* are designed to address the high estimation uncertainly for small gene expression experiments. However, using this single variance term has distinct limitations. First, it ignores the fact that the contribution of the random effect often varies widely from gene to gene[17]. Using a single variance term to account for the correlation between samples from the same individuals over-corrects for this correlation for some genes and under-corrects for others. In addition, it is a two step approach that first estimates the variance term and then estimates the regression coefficients. Thus, it does not account for the statistical uncertainty in the estimate of τ^2. Finally, it does not account for the fact that estimating the variance component changes the null distribution of $\hat{\beta}_g$. Specifically, estimating variance components in a linear mixed model can substantially change the degrees of freedom of the null distribution for fixed effect coefficients[19,20,35–37]. Ignoring this issue can lead to false positive differentially expressed genes.

Dream model

The dream model extends the previous model in order to

- enable multiple random effects
- enable the variance terms to vary across genes
- estimate residual degrees of freedom for each model from the data in order to reduce false positives
- perform hypothesis testing with moderated t-statistics using empirical Bayes approach

The definition of the dream model follows directly from the definition of the previous models. First, consider a linear mixed model for gene g with an arbitrary number of random effects:

$$\begin{aligned} y_g &= X\beta_g + \sum_j Z_j \alpha_g^{(j)} + \varepsilon_g & (11) \\ \alpha_g^{(j)} &= \mathcal{N}(0, \tau^2_{g,j}) & (12) \end{aligned}$$

where Z_j is the design matrix for the j^{th} random effect, with coefficients $\alpha_g^{(j)}$ drawn from a normal distribution with variance $\tau^2_{g,j}$. As before, heteroskedastic errors are modeled with precision weights with

$$\varepsilon_g = \mathcal{N}(0, diag(w_g)\sigma^2_\varepsilon). \tag{13}$$

In this case, estimates of coefficients $\hat{\beta}_g$ and variance components $\hat{\sigma}^2_{g,j}$ must be obtained via an iterative optimization algorithm[18].

For the linear model and generalized least squares model described above, the residual degrees of freedom is fixed based on the number of covariates and the sample size:

$$d = N - p \tag{14}$$

where N is the number of samples, and p is the number of covariates. For the single-step linear mixed model, we can explicitly account for the fact that estimating the random effect changes the residual degrees of freedom of the model[19,20,35–37]. Thus let d_g be the residual degrees of freedom for gene g. We omit the statistical details here, but d_g can be estimated from the model fit using the Satterthwaite approximation[19,37] or the Kenward-Roger approximation[20,36].

Afterwards, the moderated t-statistics are computed as

$$\tilde{t}_{g,k} \quad = \quad f_{d_0,\sigma_0^2}\left(\hat{\beta}_{g,k}, \hat{\sigma}_g^2, V_g, d_g\right) \tag{15}$$

so that $\tilde{t}_{g,k}$ follows a t-distribution with d_g+d_0 degrees of freedom. Since d_g can vary substantially across genes, the moderated t-statistics for all genes are transformed to have the same degrees of freedom. The $\tilde{t}_{g,k}$ values are transformed using the cumulative distribution function of the t-distribution to produce $\dot{t}_{g,k}$ which follows a t-distribution with d degrees of freedom. Thus the pair $\{\tilde{t}_{g,k}, d_g\}$ gives the same p-value as $\{\dot{t}_{g,k}, d\}$, but $\dot{t}_{g,k}$ is easier to interpret across genes since the transformation makes the degrees of freedom constant.

Software

The `dream()` method is available in the variancePartition[17] package (`http://bioconductor.org/packages/variancePartition/`) from Bioconductor version ≥ 3.7.

Implementation

Precision weights are estimated using `voom()` in the limma package[13,21]. Linear mixed models are estimated using `lmer()` from the lme4 package[18]. Estimating the residual degrees of freedom is performed with either Satterthwaite approximation[37] in the lmerTest package[19] or the Kenward-Roger approximation[36] in the pbkrtest package[20]. Moderated t-statistics are computed using an extension of the `eBayes()` function in the limma package[21]. Parallel processing of thousands of genes on a multi-core computer is performed with doParallel[38] and foreach[39]. Visualization is performed with ggplot2[40].

Simulations

The true expression values were simulated from a linear mixed model with two components: variance across individuals and variance across two disease classes (i.e. cases versus controls). Multiple samples from the same individual are always in the same disease class. Variance across individuals (i.e. heritability) accounts for 40% of the expression variation. Expression values were simulated for 20,738 protein coding genes from GENCODE v19, and 500 genes were differentially expressed with a fold change of of 3, corresponding to a $\log_2$ fold change of 1.58. Simulated read counts following a realistic error model were then generated using polyester v1.14.0[41]. Approximately 30 million total reads counts were generated for each sample. Simulations using a range of values for each of these parameters did not change the conclusions. Simulation and data analysis code is available at https://github.com/GabrielHoffman/dream_analysis

Data analysis

Data for Timothy syndrome[26] was downloaded from GEO at GSE25542. Data for childhood onset schizophrenia[7] was downloaded from https://www.synapse.org/#!Synapse:syn9907463. Post mortem brain RNA-seq data from Alzheimer's and controls[25] was downloaded from https://www.synapse.org/#!Synapse:syn3159438. Analysis was performed on individuals from European ancestry that were assayed in each of 4 brain regions (Brodmann areas 10, 22, 36 and 4), had ApoE genotype data, had Braak stage information, and were either controls or definite AD patients (i.e. possible and probable cases were excluded). Differential expression analysis corrected for batch, sex, RIN, rRNA rate, post mortem interval, mapping rate and ApoE genotype.

Enrichment analysis was performed with cameraPR in the limma package[23]. In order to avoid using arbitrary cutoffs to identify differentially expressed genes, gene set enrichments were evaluated by applying cameraPR to the differential expression t-statistics from each analysis.

Reproducible analysis code, figures, and statistics from differential expression and enrichment analyses are available at https://github.com/GabrielHoffman/dream_analysis.

References

[1] Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

[2] Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–20 (2013).

[3] Van Der Wijst, M. G. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* **50**, 493–497 (2018).

[4] Watson, C. T. *et al.* Integrative transcriptomic analysis reveals key drivers of acute peanut allergic reactions. *Nature Communications* **8**, 1943 (2017).

[5] Breen, M. S. *et al.* Gene networks specific for innate immunity define post-traumatic stress disorder. *Molecular Psychiatry* **20**, 1538–1545 (2015).

[6] Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics* **50**, 1–8 (2018).

[7] Hoffman, G. E. *et al.* Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. *Nature Communications* **8**, 2225 (2017).

[8] Carcamo-Orive, I. *et al.* Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. *Cell Stem Cell* **20**, 518–532.e9 (2017).

[9] Adamo, A. *et al.* 7Q11.23 Dosage-Dependent Dysregulation in Human Pluripotent Stem Cells Affects Transcriptional Programs in Disease-Relevant Lineages. *Nature Genetics* **47**, 132–141 (2015).

[10] Schwartzentruber, J. *et al.* Molecular and functional variation in iPSC-derived sensory neurons. *Nature Genetics* **50**, 54–61 (2018).

[11] Jostins, L., Pickrell, J. K., MacArthur, D. G. & Barrett, J. C. Misuse of hierarchical linear models overstates the significance of a reported association between OXTR and prosociality. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E1048 (2012).

[12] Germain, P. L. & Testa, G. Taming Human Genetic Variability: Transcriptomic Meta-Analysis Guides the Experimental Design and Interpretation of iPSC-Based Disease Modeling. *Stem Cell Reports* **8**, 1784–1796 (2017).

[13] Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, R29 (2014).

[14] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15** (2014).

[15] Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods* **14**, 687–690 (2017).

[16] Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016).

[17] Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

[18] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67** (2015).

[19] Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82** (2017).

[20] Halekoh, U. & Højsgaard, S. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software* **59**, 3–4 (2014).

[21] Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article3 (2004).

[22] Liu, R. *et al.* Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic acids research* **43**, e97 (2015).

[23] Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015).

[24] Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452–1458 (2013).

[25] Wang, M. *et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Scientific data* **5**, 180185 (2018).

[26] Pasca, S. P. *et al.* Using iPSC-derived neurons to uncover cellular phenotypes associated with Timothy syndrome. *Nature medicine* **17**, 1657–62 (2011).

[27] Ahn, K. *et al.* High rate of disease-related copy number variations in childhood onset schizophrenia. *Molecular Psychiatry* **19**, 568–572 (2014).

[28] Ahn, K., An, S. S., Shugart, Y. Y. & Rapoport, J. L. Common polygenic variation and risk for childhood-onset schizophrenia. *Molecular Psychiatry* **21**, 94–96 (2016).

[29] Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

[30] Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016).

[31] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).

[32] Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

[33] Pinheiro, J. & Bates, D. *Mixed-effects models in S and S-Plus* (Springer, New York, 2000).

[34] Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38** (1982).

[35] Hoffman, G. E. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE* **8**, e75707 (2013).

[36] Kenward, M. G. & Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–97 (1997).

[37] Giesbrecht, F. G. & Burns, J. C. Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics* **41**, 477 (1985).

[38] Weston, S. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. *R package version 1.0.11* (2017).

[39] Weston, S. foreach: Provides Foreach Looping Construct for R. *R package version 1.4.4.* (2017).

[40] Wickham, H. *Elegant Graphics for Data Analysis* (Springer, New York, 2016).

[41] Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778–2784 (2015).

Figures

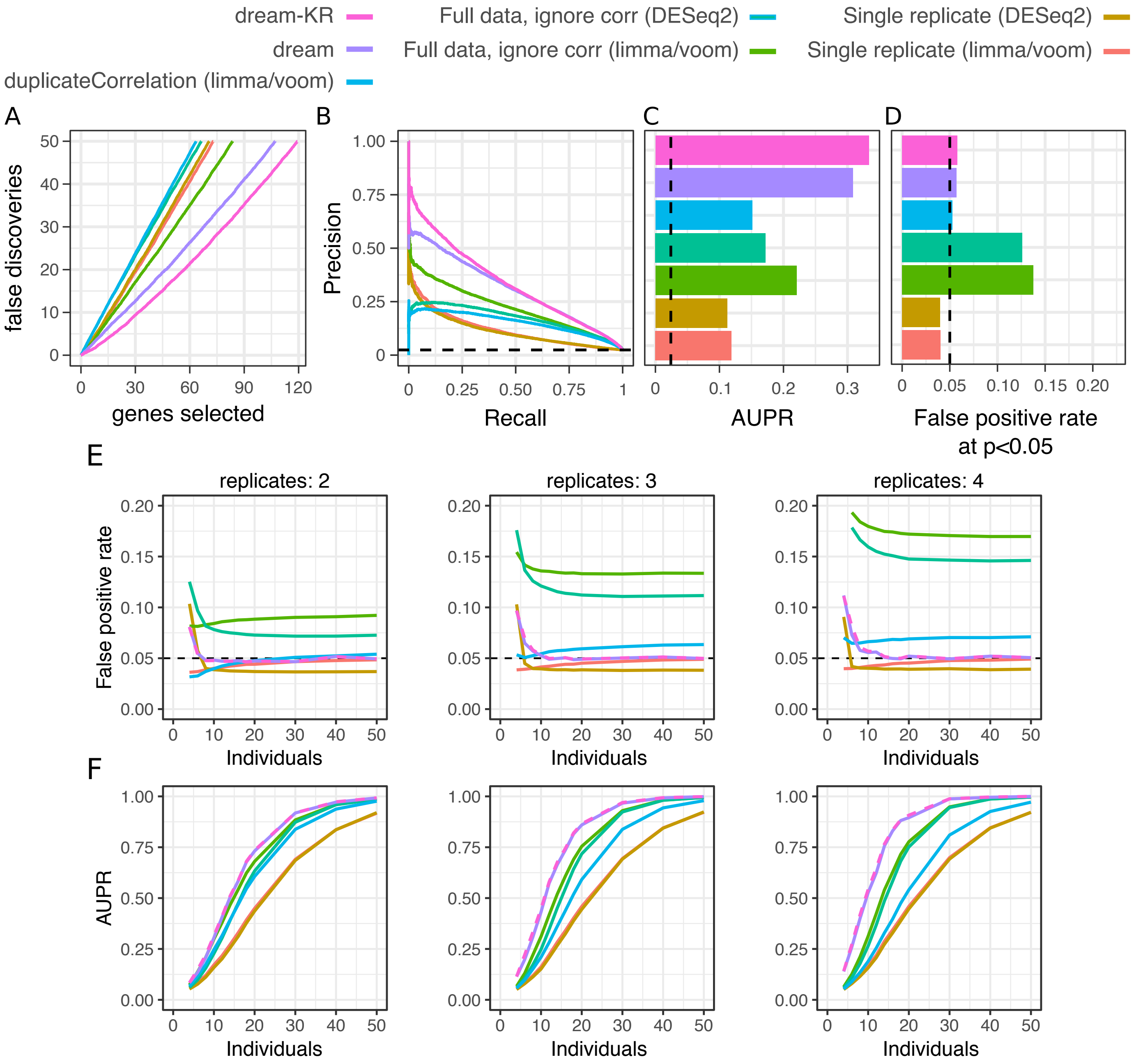


Figure 1: Performance on biologically realistic simulated data. A,B,C,D) Performance from 50 simulations of RNA-seq datasets of 8 individuals each with 3 replicates. **A)** False discoveries plotted against the number of genes called differentially expressed by each method. **B)** Precision-recall curve showing performance in identifying true differentially expressed genes. Dashed line indicates performance of a random classifier. **C)** Area under the precision-recall (AUPR) curves from **(B)**. Dashed line indicates AUPR of a random classifier. **D)** False positive rate at $p < 0.05$ evaluated under a null model were no genes are differentially expressed illustrates calibration of type I error from each method. As indicated by the dashed line, a well calibrated method should give p-values < 0.05 for 5% of tests under a null model. **E,F)** Performance summary for simulations with between 4 to 50 individuals with between 2 to 4 replicates. For each condition, 50 simulations were performed for a total of 1800. **E)** False positive rate at $p < 0.05$ for simulations versus the number of individuals and replicates. Black dashed line indicates target type I error rate of 0.05. **F)** AUPR for simulations versus the number of individuals and replicates.

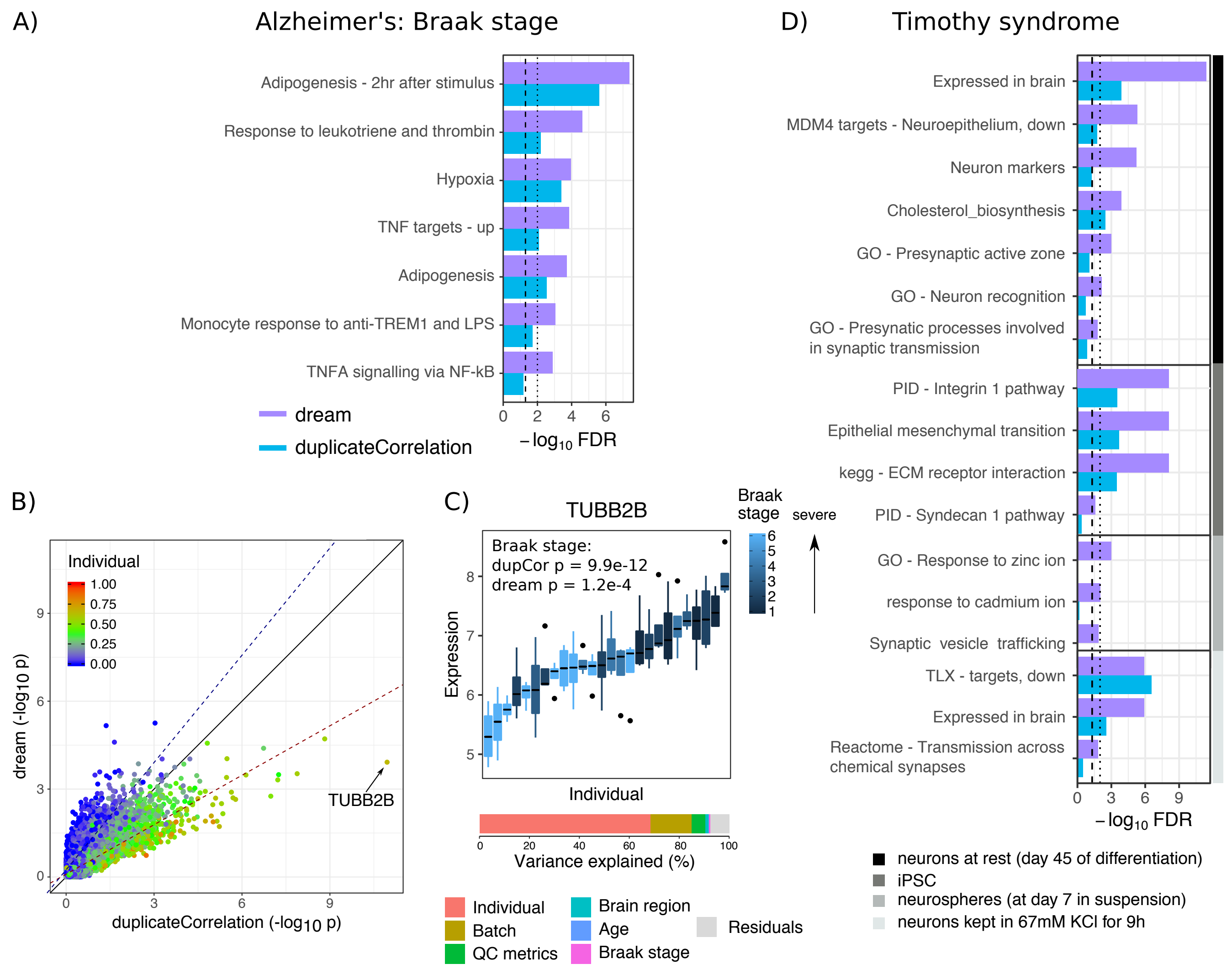
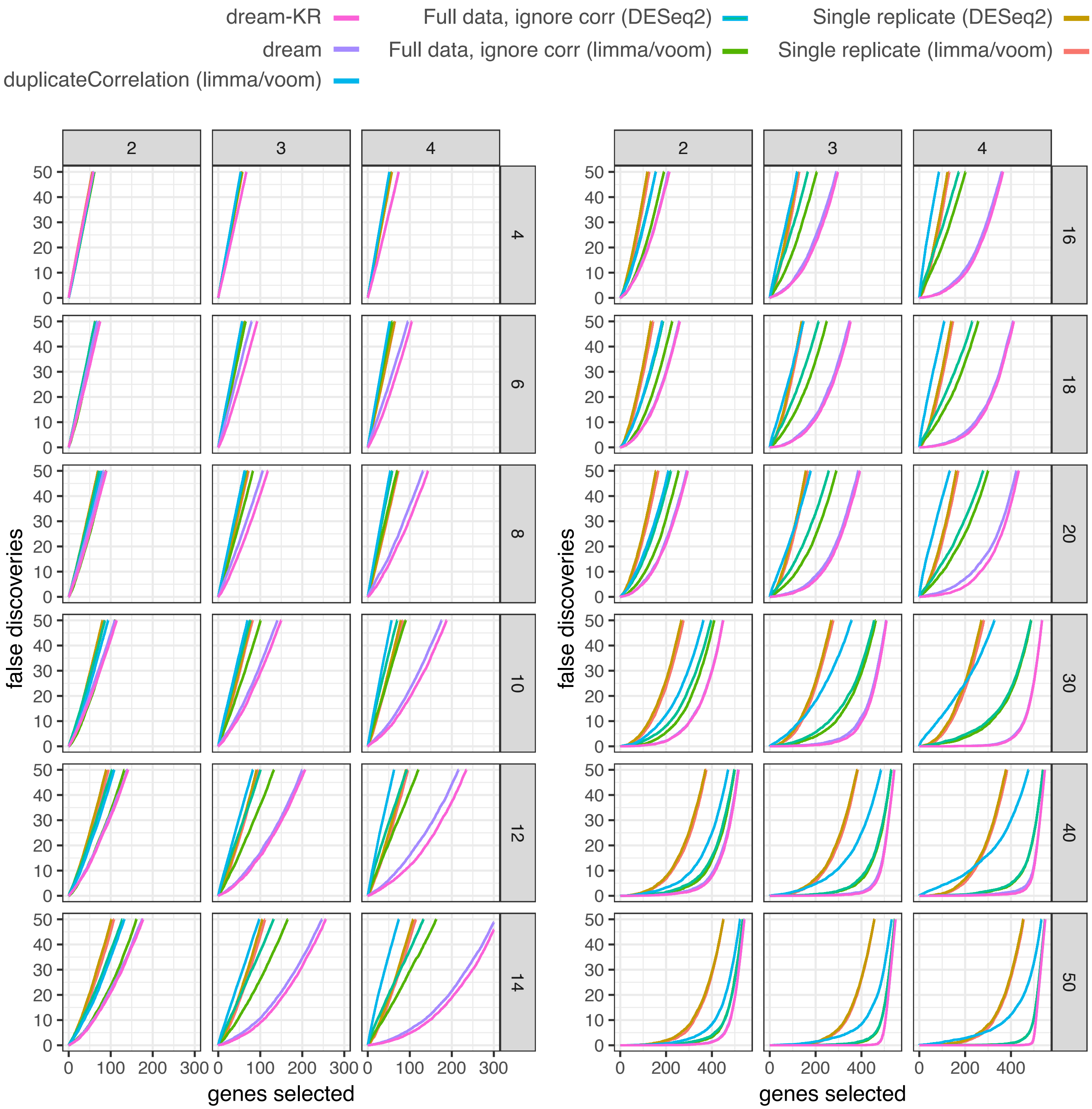
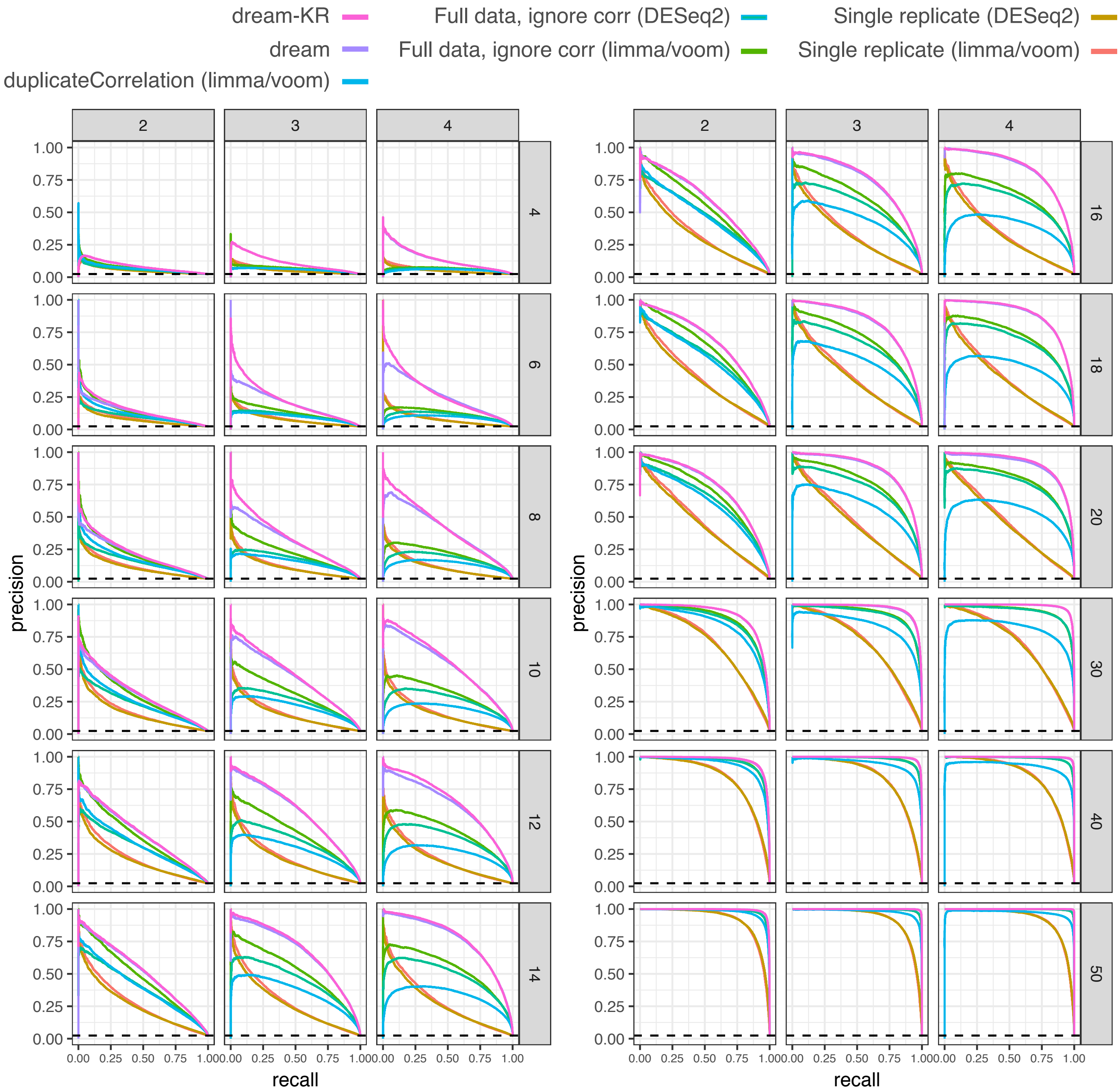


Figure 2: Application to transcriptome data from neuropsychiatric disease. A) Gene set enrichment FDR for genes associated with Braak stage. Results are shown for dream and duplicateCorrelation. Lines with broad and narrow dashes indicate 5% and 1% FDR cutoff, respectively. **B**) Comparison of $-\log_{10}$ p-values from applying dream and duplicateCorrelation to Braak stage. Each point is a gene, and is colored by the fraction of expression variation explained by variance across individuals. Black solid line indicates a slope of 1. Dashed line indicates the best fit line for the 20% of genes with the highest (red) and lowest (blue) expression variation explained by variance across individuals. **C**) Expression of TUBB2B stratified by individual and colored by Braak stage so that each box represents the expression in the multiple samples from a given individual. Bar plot of variance decomposition shows that 68.3% of variance is explained by expression variance across individuals. Since this value is much larger than the genome-wide mean, duplicateCorrelation under-corrects for the repeated measures. **D**) Gene set enrichment FDR for genes associated with Timothy syndrome compared to controls in four cell types or conditions.

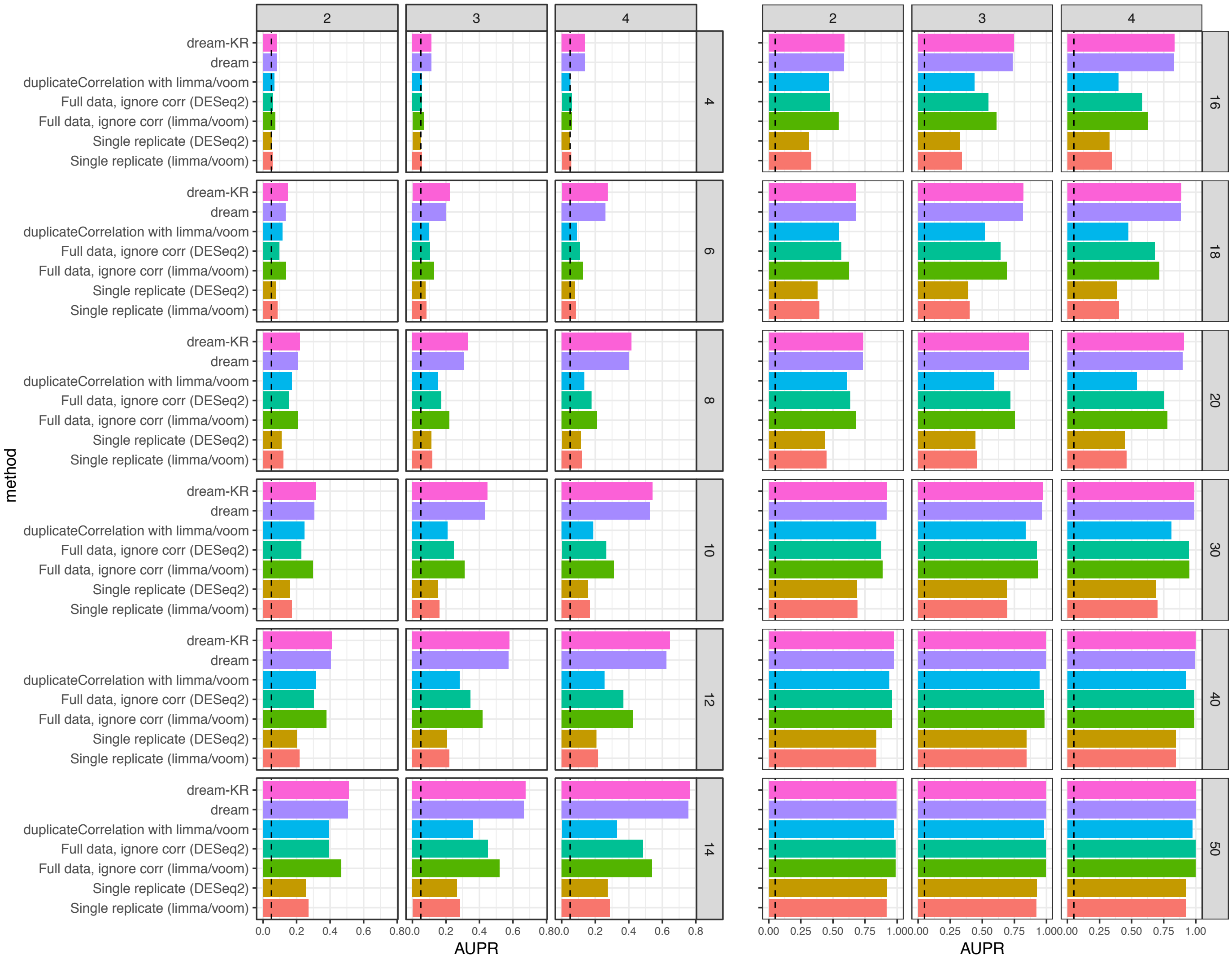
Supplementary Figures



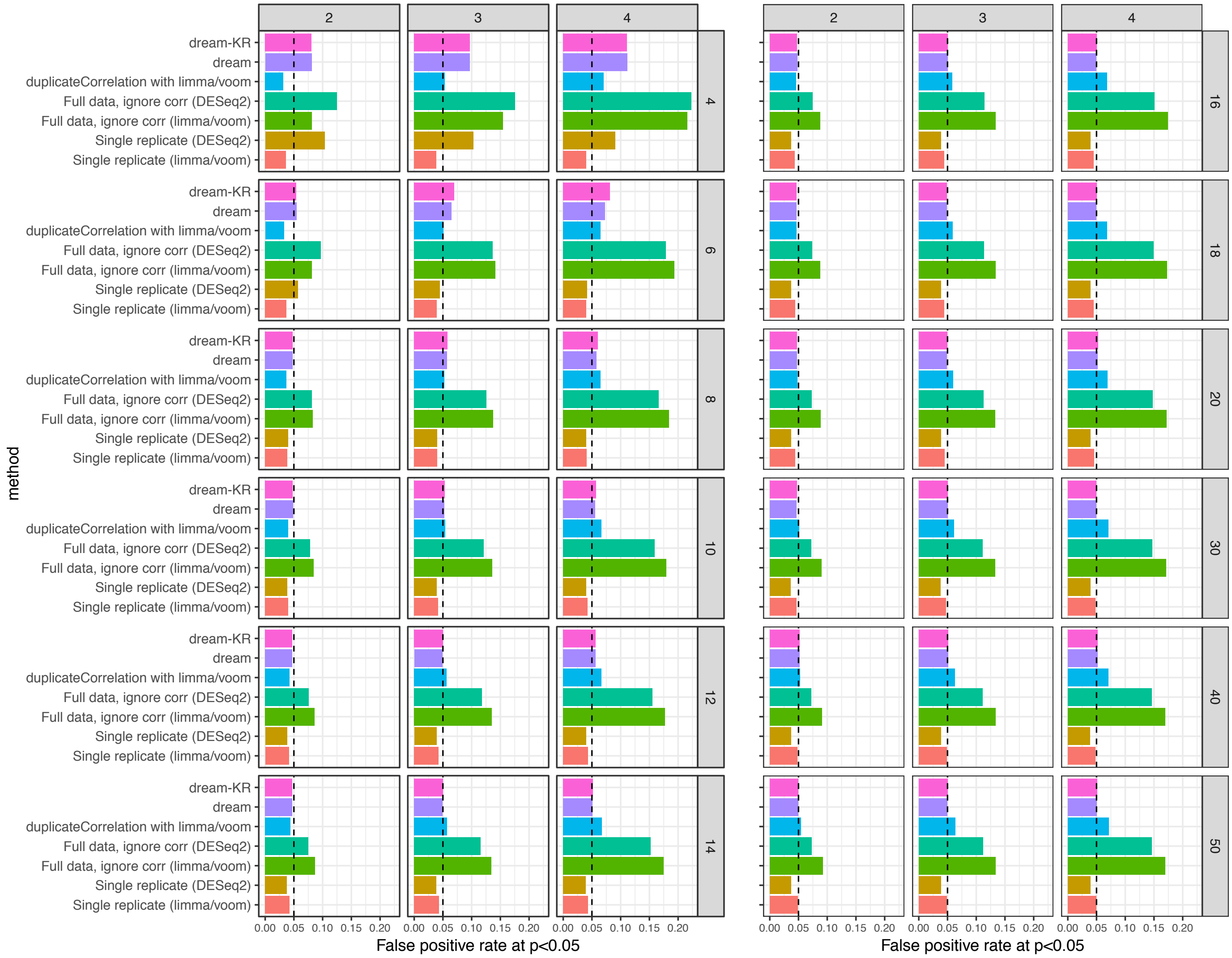
Supplementary Figure 1: False discovery rates for multiple simulation conditions. False discoveries plotted against the number of genes called differentially expressed by each method. Results are shown for between 4 and 50 individuals (rows) and 2 to 4 replicates (columns). For each combination, 50 simulated datasets were analyzed.



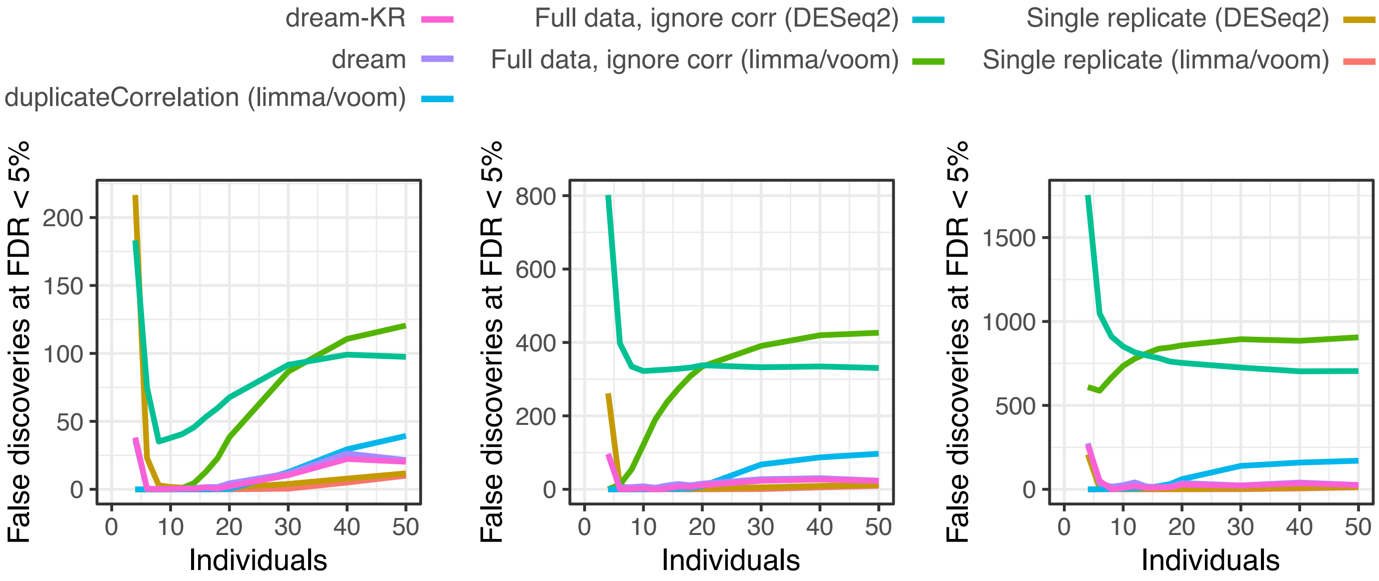
Supplementary Figure 2: Precision-recall curves for multiple simulation conditions. Plots shows performance in identifying true differentially expressed genes. Dashed lined indicates performance of a random classifier. Results are shown for between 4 and 50 individuals (rows) and 2 to 4 replicates (columns). For each combination, 50 simulated datasets were analyzed.



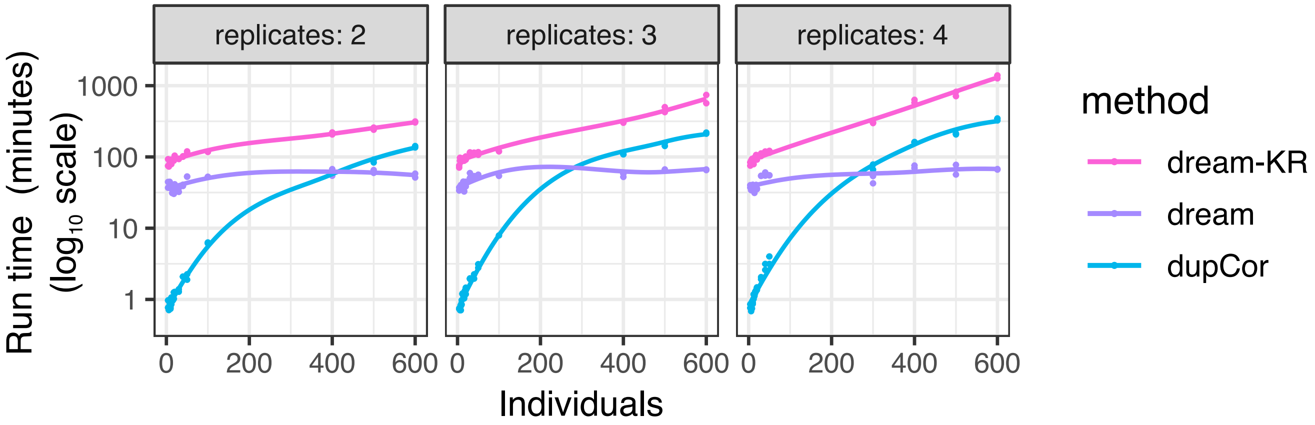
Supplementary Figure 3: Area under the precision-recall (AUPR) for multiple simulation conditions. Dashed line indicates AUPR of a random classifier. Results are shown for between 4 and 50 individuals (rows) and 2 to 4 replicates (columns). For each combination, 50 simulated datasets were analyzed.



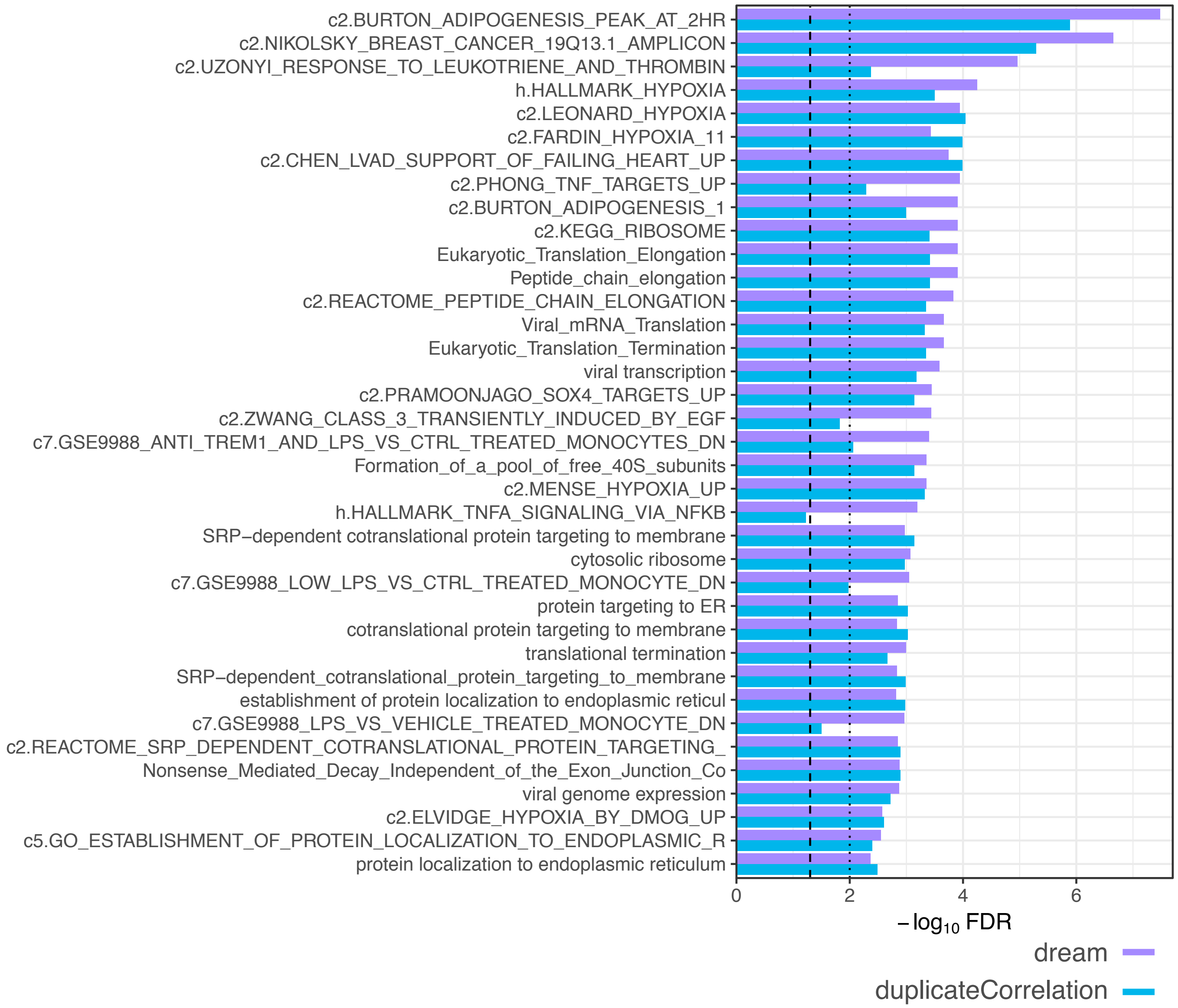
Supplementary Figure 4: False positive rate for multiple simulation conditions. False positive rate at p $<$ 0.05 evaluated under a null model were no genes are differentially expressed illustrates calibration of type I error from each method. As indicated by the dashed line, a well calibrated method should give p-values $<$ 0.05 for 5% of tests under a null model. Results are shown for number of individuals between 4 and 50 (rows), and replicates between 2 and 4 (columns).



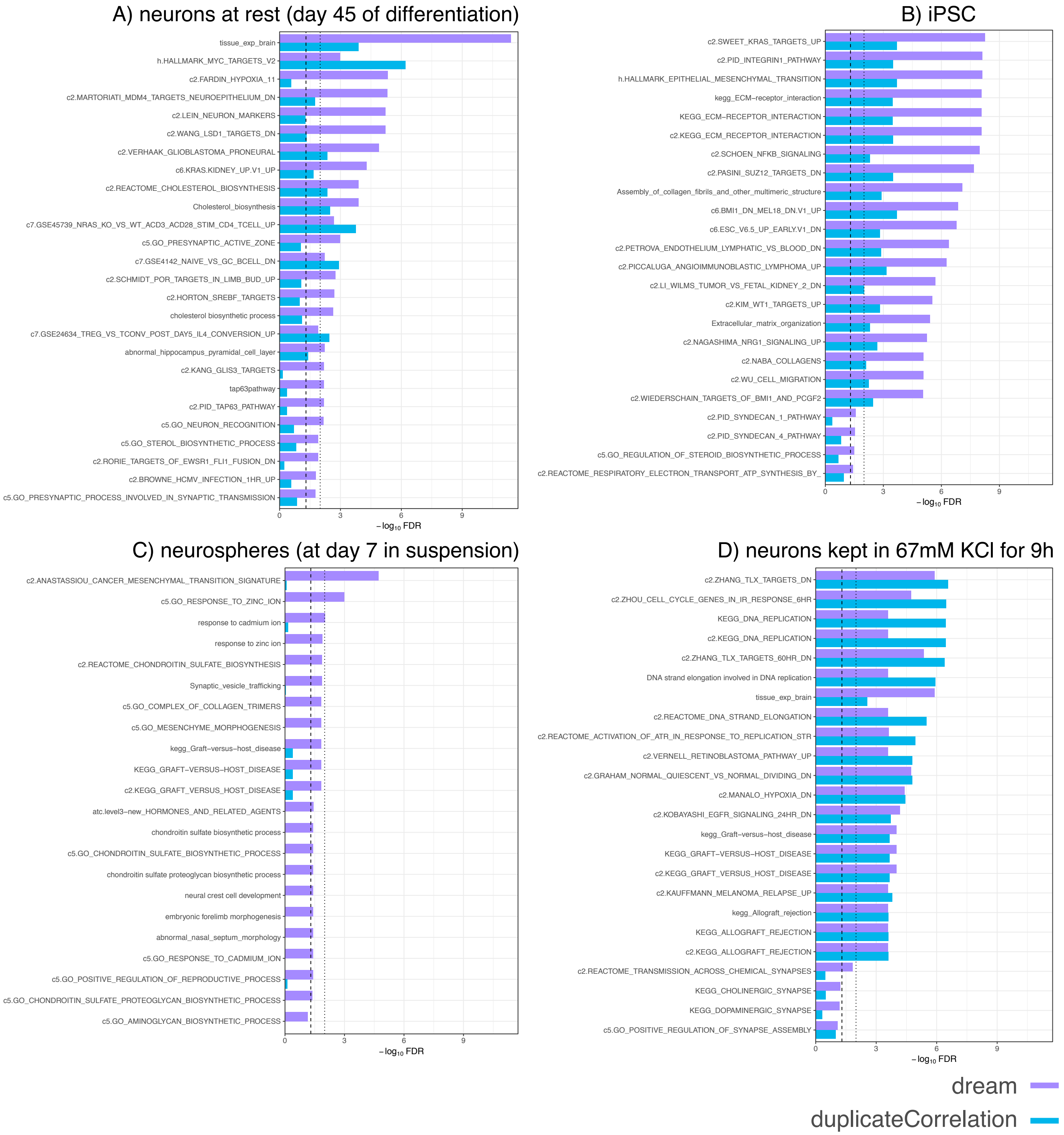
Supplementary Figure 5: Number of genes passing FDR cutoff of 5% in a null simulation. False discoveries plotted against number of individuals for simulations with between 4 to 50 individuals with between 2 to 4 replicates from Figure 1. The values shown are averaged across 50 simulations for each condition. The analysis considered only genes that were not differentially expressed in the simulation, so that there are no true positive genes and all positive genes are false positives.



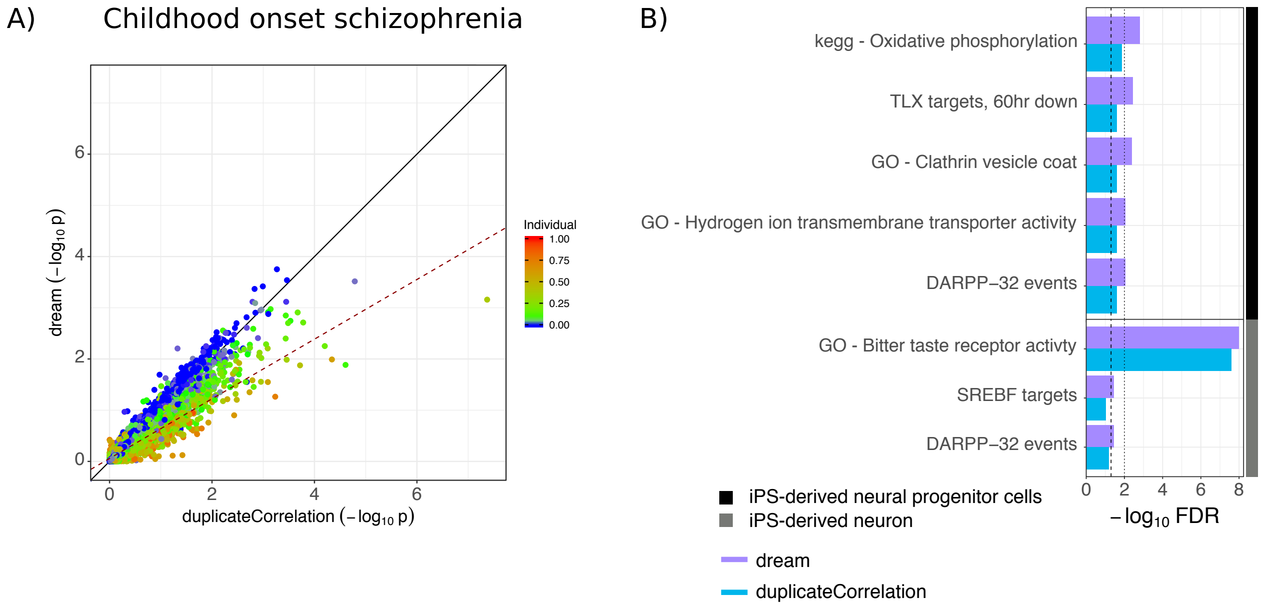
Supplementary Figure 6: Run time comparison. Run time for was evaluated on the simulated datasets using 12 threads on a 12 core Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz. Each combination of individuals, replicates, methods and threads was evaluated on 2 simulated datasets. Lines show loess smoothing. The formula used was: ∼ `Disease + (1|Individual)`.



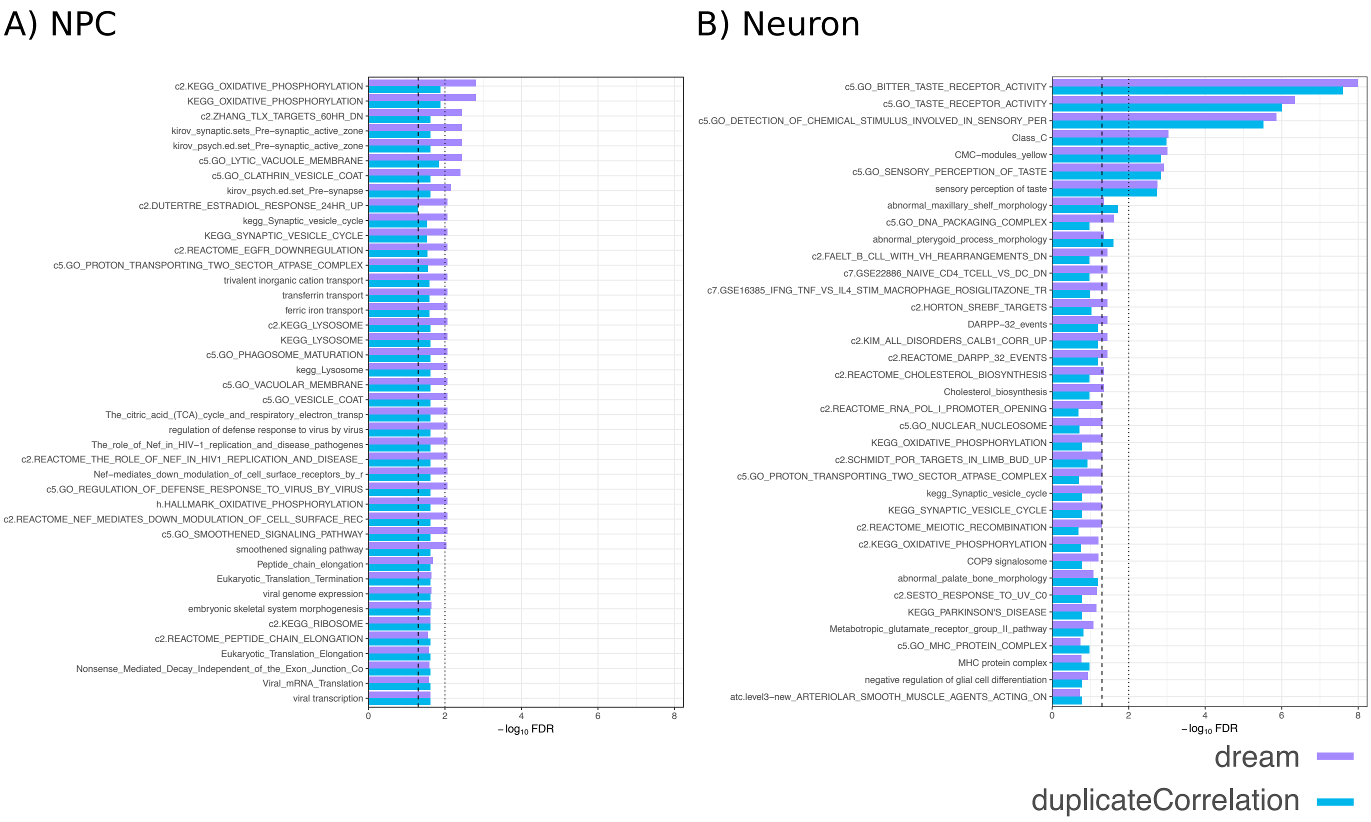
Supplementary Figure 7: Gene set enrichment FDR for top 30 genesets from differential expression analysis of Braak stage. Enrichment FDRs were computed using t-statistics from dream and duplicateCorrelation.



Supplementary Figure 8: Gene set enrichment FDR for top 20 genesets from differential expression analysis of Timothy Syndrome. Enrichment FDRs were computed using t-statistics from dream and duplicateCorrelation analysis of **A**) neurons at rest, **B**) iPSC, **C**) neurospheres, and **D**) neurons in 67 mM KCl for 9 h.



Supplementary Figure 9: Application to transcriptome data from childhood onset schizophrenia. **A**) Comparison of $-\log_{10}$ p-values from applying dream and duplicateCorrelation to disease status in neurons. Each point is a gene, and is colored by the fraction of expression variation explained by variance across individuals. Black solid line indicates a slope of 1. Dashed line indicates the best fit line for the 20% of genes with the highest (red) and lowest (blue) expression variation explained by variance across individuals. **B**) Gene set enrichment FDR for genes associated with disease status in iPS-derived neurons and neural progenitor cells. Results are shown for dream and duplicateCorrelation. Lines with broad and narrow dashes indicate 5% and 1% FDR cutoff, respectively.



Supplementary Figure 10: Gene set enrichment FDR for top 30 genesets from differential expression analysis of childhood onset schizophrenia. Enrichment FDRs were computed using t-statistics from dream and duplicateCorrelation analysis of iPSC-derived **A**) neural progenitor cells (NPCs) and **B**) neurons.